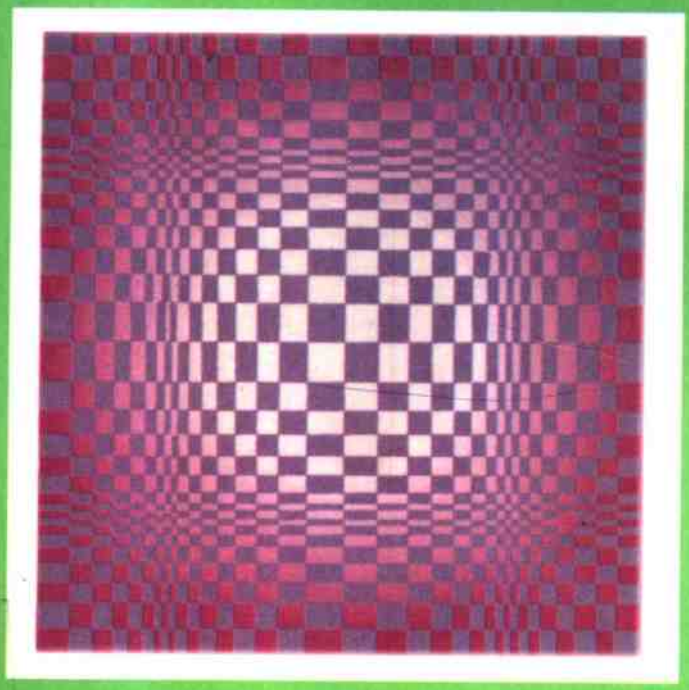
科学数据库与信息技术论文集

第五集

中国科学院科学数据库中心　编



中国科学技术出版社

科学数据库与信息技术
论　文　集

第五集

中国科学院科学数据库中心　编

图书在版编目(CIP)数据

科学数据库与信息技术论文集．第5集/中国科学院科学数据库中心编．—北京：中国科学技术出版社，2000.9

ISBN 7-5046-2911-1

Ⅰ.科… Ⅱ.中… Ⅲ.①科学技术-专用数据-数据库-文集②信息技术-文集 Ⅳ.TP392-53

中国版本图书馆CIP数据核字(2000)第67363号

中国科学技术出版社出版

北京海淀区白石桥路32号 邮政编码：100081

电话：62179148 62173865

新华书店北京发行所发行 各地新华书店经售

北京地质印刷厂印刷

*

开本：787毫米×1092毫米 1/16 印张：17.625 字数：405千字

2000年9月第1版 2000年9月第1次印刷

印数：1— 500册 定价：35.00元

内容简介

本书共收入论文45篇。这些论文主要反映我国近年来科学数据库在建库技术、网络技术、信息服务、总体发展等方面所取得的成果，也反映了围绕着这些问题在学术上取得的进展。集中体现了近年来国内数据库与信息技术的研究和应用水平及发展历程。

本书可供从事数据库技术、网络技术和信息系统研究的科技人员、工程技术人员，以及相关学科的研究人员、大专院校师生参考。

责任编辑　屈惠英
责任校对　冯　静
责任印制　王　沛

序　言

科学数据库及其信息系统经过近 20 年的建设与发展，形成了一个从基本的数据库检索直至专业咨询、决策服务的完整的科技信息服务系统。目前已建成了分布与集中相结合的上百个各种类型的专业数据库，总数据量达 5 600 亿字节(560GB)。其中，2 000 亿字节(200GB)数据已上网，由中心站点和分布在网上本地和外地的相互独立的若干个专业库子站点组成了网上的科技信息服务体系，向国内外用户提供服务。同时，这批宝贵的信息资源已在国家经济建设、国防建设、规划决策、科学研究、科技攻关、学科发展、国际合作等诸多方面得到应用，取得了显著的社会效益和一定的经济效益，在国内外产生了一定影响。为此，科学数据库及其信息系统于 1997 年获中国科学院科技进步一等奖，1998 年获国家科技进步二等奖。

十多年来，随着计算机技术的进步和发展，科学数据库系统也不断改进和完善。科学数据库的研制过程正是知识积累、加工、利用和传播的过程，是跟踪数据库技术、网络技术发展的过程。每两年召开一次的科学数据库与信息技术学术讨论会是科学数据库的建设者进行工作交流、学术研讨和成果展示的场所，集中展现了国内数据库与信息技术的发展历程。

2000 年 10 月，在中国上海召开的第五届“科学数据库与信息技术”学术讨论会，主要是进一步研讨自 1998 年第四届学术讨论会以来，数据库建库及管理技术、网络技术、信息服务技术，以及对新技术应用的探讨等方面的进展和经验，反映在建库技术、应用服务、总体发展等方面取得的成果。这次会议收到的论文，经编辑委员会认真评审，选出了 45 篇进行会议交流并收入了本次会议的论文集。这些论文集中体现了近年来国内数据库与信息技术的研究和应用水平，可以提供从事数据库及信息系统研究的同行参考。

希望科学数据库在今后的发展过程中，结合国家的创新体制，配合中国科学院知识创新工程中基础研究、战略性研究，以及知识传播和科学普及等，发挥出其丰富的科技信息资源的巨大优势，为我国的经济建设，特别是高科技的发展作出更大的贡献。

中国科学院科学数据库专家委员会主任

2000 年 3 月 28 日

目　录

文献数据库及全文检索技术

数据库应用

总论

基于科学数据库科研环境建设的总体设想

李望平　阎保平　孙九林　罗晓沛
（中国科学院科学数据库，北京 100080）

摘要　本文介绍了基于科学数据库科研环境建设的总体设想，包括其指导思想、建设目标和建设内容等，阐述了科学数据库在今后一段时期内的发展方向。

关键词　科学数据库　科研环境　建设　总体设想

一、引　言

20 世纪 50～60 年代，由于计算技术的发展和成熟，使大量数据的收集、加工、存储和利用成为可能，致使数据成为可能产生经济和社会效益的重要资源；70 年代以来，计算机软件和硬件技术的发展，使对大量数据的精细加工，使数据变成信息并加以利用成为可能；当前，由于计算机技术和通信技术的发展，计算机与网络的密切结合，使信息的传播和利用超脱了时空的限制，成为社会发展和进步的极为重要的、可共享的资源。信息的来源大多为利用知识工具对数据的深层加工，科学数据库新积累的数据必须在经过加工后才能升华为有用信息，这将形成科学数据库系统今后的重要发展方向。丰富的数据，完善的加工工具，将为知识创新工程提供支持，这正是基于科学数据库科研环境建设的目的所在。

二、系统基础

中国科学院科学数据库的发展过程正是利用计算机技术进行知识积累、加工和利用的过程。中国科学院作为中国自然科学的研究中心，在长期的科学研究实践中，通过观测、考察、试验、计算等多种途径产生和积累了大量具有重要科学价值和实用意义的科学数据和资料。20 世纪 60 年代发展起来的数据库技术，为有效管理和开发应用科学数据

创造了有利条件，70 年代开始各研究所在自己学科领域中试建数据库，专业库的建立促进了数据科学管理的进程，然而分散的、独立的库，限制了数据规范标准的统一和共享的实现。为此，中国科学院 1982 年提出了“科学数据库及其信息系统”的建设项目，随后的十几年中，科学数据库先后得到了国家计委、中国科学院和国家基金委的立项支持，使系统不断取得突破性进展。十多年来，随着计算机技术的进步和用户需求的不断变化，系统也不断改进和完善。科学数据库的建设过程正是知识积累、加工、利用和传播的过程，是跟踪数据库技术、网络技术发展和应用的过程，也是广大系统建设者与用户不断沟通、向实用化发展的过程。为此，科学数据库及其信息系统于 1997 年获中国科学院科技进步一等奖，1998 年获国家科技进步二等奖。

科学数据库采用逻辑上集中、物理上分散的建库和运行服务体系，形成了一个从基本的数据库检索直至专业咨询、决策服务的完整的科技信息服务系统。现在科学数据库已有专业建库单位 21 个，专业数据库 126 个，总数据量达 5 600 亿字节(560GB)；科学数据库基于中国科技网对国内外用户提供服务，已在中国科技网上建立集中与分布的 Web 站点 15 个，上网专业数据库 100 个，数据量约 2 000 亿字节(200GB)。科学数据库由中心站点和分布在网上本地和外地的相互独立的若干个专业库子站点组成了网上的科技信息服务体系。

科学数据库经过十几年的建设与发展，已经成为目前国内信息量最大，学科专业最广，服务层次最高，综合性最强的科学信息服务系统。当前的外部条件也给科学数据库带来了进一步发展的最佳时机。

“中关村科技园区信息化建设”经国务院批准，已经开始启动。该工程建设中的网络改造将以“中国科技网”为基础，在中关村地区建成 1 000Mbps 的高速光纤区域网。中国科技网的进一步建设发展，无疑为作为“中国科技网”上重要科技信息资源的“科学数据库”创造了优越的网络环境。

中国科学院知识创新工程于 2000 年在计算机网络信息中心安装曙光 2000 超级计算机，使中国科技网上的超级计算能力达到浮点运算速度每秒 1 200 亿次以上。这将对科学数据库的在知识创新工程中的应用提供了更加广泛的支持。

从国际国内网络信息服务业的发展趋势来看，由于网络环境的统一，ISP 相对整合，ICP 物理上分散。二者的发展趋势正迎合了科学数据库多年来行之有效的运行服务体系。

三、 建设目标

随着知识经济时代的到来，现在的国际竞争已经由过去争夺自然资源发展到争夺信息资源，争夺人才，争夺知识，争夺高技术制高点。科学数据库是中国科学院十几年来科研综合信息和知识的积累和应用，正迎合了我国知识经济时代到来的时机，顺应了中国科学院建设国家知识创新系统和知识创新基地的需求。

知识创新的动力来源于知识的积累和发展，科学数据资源，即科学数据、科学事实、科学文献、科学思想是科学数据库的收集和存储对象。有效地利用已有大量的、多种学科的信息资源的关键是提供较完善的对科学资源收集、整理、存储、加工和传播的工具，从而提供完善的研究环境。现代信息技术的发展，特别是计算机硬件、软件和网络技术的发展为

科学资源的有效利用提供了极其有利的条件。并认为，知识创新的基础和前提应是思维方法、研究工具和科研环境的创新。

科学数据库今后发展的总体目标是建立基于科学数据资源，并提供现代信息技术（处理）手段，创建新的科学研究环境，从而有效地改善知识创新工程实施的条件和环境。

基于科学数据库的信息化的科研环境的建立正是为实现这种目标所采取的步骤，基于计算机数据库、网络互联环境，可使科学和工程数据的传播和利用方式得到改变，可以为科研人员提供高效率、高质量的服务。同时，提供对数据的加工工具和完善的工作环境，则形成一种全新的科研和开发方式。

四、 建设内容

基于科学数据库的信息化科研环境建设将要进行以下三个方面的工作。

1. 科学数据资源的收集和积累

（1）实验室产生的动态数据积累。根据科学研究的特点，对实验室里在实验过程中产生的数据，以及分析测试数据进行收集、整理、加工和建库保存，提供资源共享。

（2）专业领域科研数据的积累与保存。将科研课题、项目的研制过程中采集和产生的数据资料积累保存。信息和数据的收集、分类、存储、检索、加工和避免重复查询等。

（3）国外科学数据的合作交换，扩大专业领域的数据资源。通过与国外同行的合作交换，联系专业领域内国际上最权威的网上免费数据服务系统，并在我国建立相应的网上镜像（Mirror）节点。使我国科研人员可直接在国内网上检索最新国外科技信息。

（4）扩大学科专业领域的数据收集积累和共享。将科学数据库现有建库单位以外的研究所组织进来，建库上网，形成一个学科全面的科技数据库群体。

2. 科学数据资源利用环境的建设

（1）数据向信息和知识的加工转变。数据的加工与处理，根据科研的需求，把数据加工和处理成有用的信息。

（2）知识加工的软件工具开发。针对专业研究人员应用的需求，开发知识加工的工具。常规工具和工作有：全文检索工具、超文本/超媒体工具、图像识别/检索工具、多媒体管理/处理工具、辅助设计/验证工具、统计分析工具、数据挖掘/钻探工具、交互式/非交互式研讨工具、文字处理/文件生成工具、专业性智能推理软件工具等。上述有的已有商品化的产品，有的则需要针对特定的需求进行开发或再开发。

（3）构造集成化的数据库系统环境。研制集成化数据库系统和应用软件工具，开发专业领域数据库的公共界面，建立面向问题的数据库系统。建立网上专业综合科技信息站点，将自建数据库与 Internet 上收集的科技信息源综合集成。

（4）建立基于科学数据库的虚拟科研环境。这是一个计算机辅助科研工作综合应用工具和支撑环境，它的数据信息和工具信息都有一个管理和环境信息库（EIB）控制的集成化系统。通过开发面向应用、面向问题、面向行业的应用软件和专家系统，使专业领域数据与知识得到综合利用。

3. 科学数据资源服务体系的建立

(1)建立科学数据中心。上述"科学数据资源利用环境建设"对海量存储、快速查询所需的高速网络信道和网上科学超级计算能力提出了更高层次的需求。即通过建立科学数据中心,提供一个可满足上述应用需求的高层次信息技术服务平台的支撑环境来实现资源共享。

(2)开展专业信息服务。中国独有资源数据库上网服务,通过中国独有资源数据库的上网服务,与国外同行建立起互惠互利的合作关系,达到全球性科技资源共享。

(3)开展公益性科技信息服务。组织和集成 Internet 上已有的国内外科技信息资源,开展综合科技信息服务。包括科技期刊、科技新闻、科技政策、科技项目、科技人才、科技书目、学位论文、专利信息、标准信息和科技专业应用软件,以及国内外其他可以利用的信息资源,以现有成熟的技术广泛吸收和集成 Internet 业已存在的科技信息资源,使用户在网上能够及时地获取有关信息。

(4)开展科学普及服务。基于科学数据库开发建立《中国科普博览》网站,进行科学知识的传播。科普工作是科教兴国的重要组成部分,是知识创新工程中知识传播的重要内容,社会对于科普的需求十分强烈。网上科普内容的建立是一种方便、快速和广泛的新型科普形式。科学数据库现有 126 个专业数据库,随着它的发展,还会增加更多的不同专业的数据库。这些专业数据库是加工建立科普内容信息的最基本的资源。利用专业数据库中有关知识和内容,加以整理和改造,综合计算机、网络、多媒体技术和虚拟现实技术,声、图、文并茂,适合于网上不同层次的用户。

(5)开展面向问题的专业信息服务或信息推送。针对国家和院一些重大项目进行专业信息服务或信息推送。建立各专业数据咨询、服务小组。

五、结　语

以上介绍了基于科学数据库的信息化科研环境建设的几个方面的工作,作为中国科学院多学科群共享的基础环境的建设,应作为知识创新科研环境的建设在全院范围内进行组织实施。将多年积累起来的科技信息资源组织起来,专业数据库结合相应学科领域的知识创新研究,应用计算机、数据库和网络等先进技术,实现资源的共享和协同工作,建立信息化的知识创新科研环境。

基于科学数据库的信息化科研环境建设具体实施应分阶段进行。首先在短期内进行学科试点,选择科学数据库中有条件的专业数据库建库单位按学科进行知识创新科研环境的建设试点工作。随后总结试点学科的经验,基于科学数据库建设的基础,按"十五"计划立项在全院各个学科领域中全面实施。

随着信息时代的到来,科学数据库在今后的发展过程中,结合国家的创新体制,改善知识创新环境,配合中国科学院知识创新工程中基础研究、战略性研究,以及知识传播和科学普及等,必将发挥出其丰富的科技信息资源的巨大优势,做出更大的贡献。

参 考 文 献

1 李望平，孙九林，张建中．科学数据库及其信息技术论文集(第四集)．科学数据库及其信息系统的研制与应用．北京：科学出版社，1998.

2 罗晓沛．科学数据库及其信息技术论文集(第四集)．基于科学数据库的虚拟科研环境的研究．北京：科学出版社，1998.

TENTATIVE PLAN FOR OVERAL DESIGN OF RESEARCH ENVIRONMENT BASED ON SCIENTIFIC DATABASE

Li Wangping　Yan Baoping　Sun Jiulin　Luo Xiaopei

(Scientific Database of Chinese Academy of Sciences, Beijing 100080)

Abstract　This paper gives an account of the tentative plan for the overall design of the research environment based on the scientific database, which includes guiding principle, objective and contents of the database construction, thus expounding the development orientation of the Scientific Database in the coming years.

Keywords: scientific database; research environment; construction; tentative plan for overall design.

数据挖掘在科学数据库中的应用探索

罗晓沛
（中国科学技术大学研究生院，北京 100039）

摘要 本文讨论科学数据库新的应用技术，介绍数据挖掘技术的基本内容以及探索其对科学数据库应用的途径。

主题词 数据挖掘 科学数据库

一、引 言

数据挖掘的目的是从大量数据中寻找有用的信息，它起先主要应用于商业活动，例如市场管理、风险管理和欺诈管理。它能否应用于对科学数据的加工，并从已有的科学数据库中寻找出新的科学知识或规律，是本文提出的并想探讨的问题。想法是，既然可从大量的商业活动所积累的数据中挖掘出有用的信息，那么就应该有可能从大量科研活动所积累的数据中挖掘出我们还未掌握的知识，即新的科学发现。作者预测：数据挖掘技术应该成为对科学数据加工的一种新的技术，至少应该运用这种技术对大量科学数据的加工做出尝试，因此科学工作者应了解数据挖掘的技术、方法、过程和步骤，并探索其对科学数据挖掘的潜在应用或应用领域。

二、数据挖掘技术

数据挖掘是指一个完整的过程，该过程从大型数据库中挖掘先前未知的，有效的，可实用的信息，并使用这些信息做出决策或丰富知识。

数据挖掘环境可示意如下图 1：

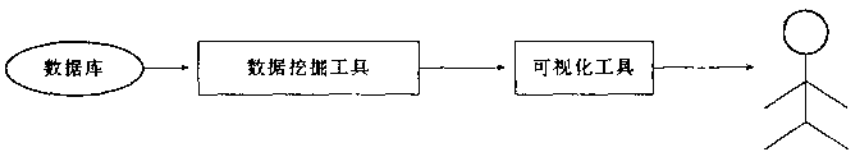


图 1 数据挖掘环境框图

数据挖掘与传统的数据分析（如查询、报表、联机应用分析）的本质区别是数据挖掘，是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘所得到的信息应具有先前未知、有效和可实用 3 个特征。

先前未知的信息是指该信息是预先未曾预料到的，即数据挖掘是要发现那些不能靠直觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。在商业应用中最典型的例子就是一家连锁店通过数据挖掘发现了小孩尿

布和啤酒之间有着惊人的联系。

信息的有效要求挖掘前要对被挖掘的数据进行仔细检查，保证它们的有效性，才能保证挖掘出来的信息的有效性。从某种程度来讲，科学数据的有效性与其他数据相比往往是能得到保证的。

最为重要的是要求所得的信息是有可实用性，即这些信息或知识对于所讨论的业务或研究领域是有效的，是有实用价值和可实现的。常识性的结论，或被人们或竞争对手早已掌握的或无法实现的事实都是没有意义的。

三、数据挖掘过程

图 2 描述了数据挖掘的基本过程和主要步骤。

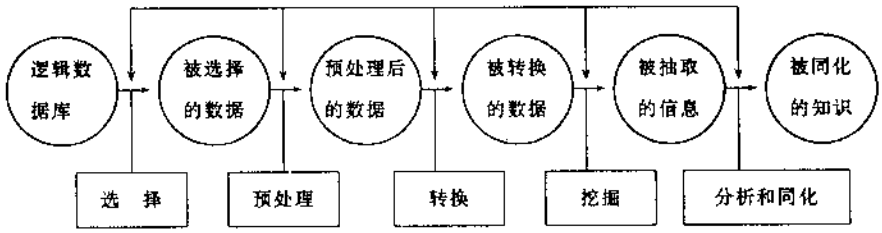


图 2 数据挖掘过程的步骤

在数据挖掘中被研究的业务对象是整个过程的基础，它驱动了整个数据挖掘过程，也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。图 2 各步骤是按一定顺序完成的，当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的，绝大多数的工作需要人工完成。图 3 给出了各步骤在整个过程中的工作量之比。可以看到，60％的时间用在数据准备上，这说明了数据挖掘对数据的严格要求，而后挖掘工作仅占总工作量的 10％。

数据挖掘过程中各步骤的大体内容如下：

1. 确定业务对象

清晰地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的，但要探索的问题应是有预见的，为了数据挖掘而数据挖掘则带有盲目性，是不会成功的。

2. 数据准备

1）数据的选择

搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据。

2）数据的预处理

研究数据的质量，为进一步的分析作准备。并确定将要进行的挖掘操作的类型。

3）数据的转换

将数据转换成一个分析模型。这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

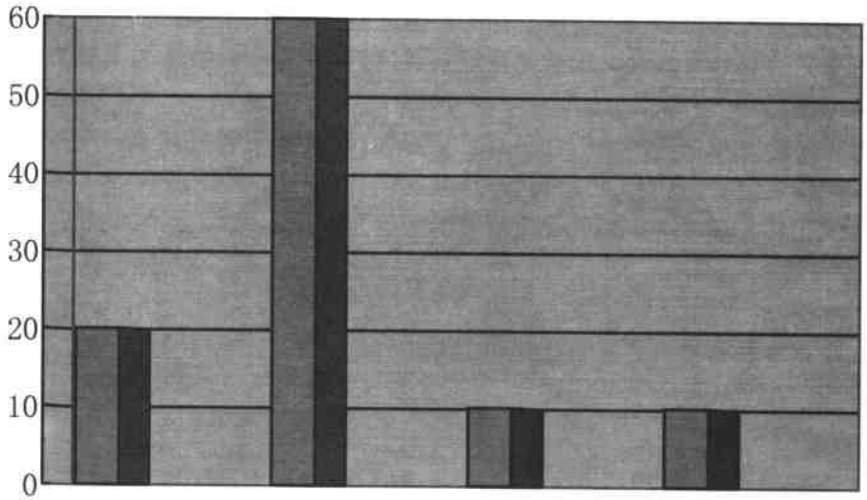


图 3　数据挖掘过程工作量比例

3. 数据挖掘

对所得到的经过转换的数据进行挖掘。除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。

4. 结果分析

解释并评估结果。其使用的分析方法一般应作数据挖掘操作而定，通常会用到可视化技术。

5. 知识的同化

将分析所得到的知识集成到业务信息系统的组织结构中去。

数据挖掘过程的分步实现，不同的步会需要是有不同专长的人员，他们大体可以分为三类。

业务分析人员：要求精通业务，能够解释业务对象，并根据各业务对象确定出用于数据定义和挖掘算法的业务需求。

数据分析人员：精通数据分析技术，并对统计学有较熟练的掌握，有能力把业务需求转化为数据挖掘的各步操作，并为每步操作选择合适的技术。

数据管理人员：精通数据管理技术，并从数据库或数据仓库中收集数据。

从上可见，数据挖掘是一个多种专家合作的过程，也是一个在资金上和技术上高投入的过程。

四、数据挖掘技术演变

数据挖掘其实是一个逐渐演变的过程，电子数据处理的初期，人们就试图通过某些方法来实现自动决策支持，当时机器学习成为人们关心的焦点。机器学习的过程就是将一些已知的并已被成功解决的问题作为范例输入计算机，机器通过学习这些范例总结并生成相应的规则，这些规则具有通用性，使用它们可以解决某一类的问题。随后，随着神经

网络技术的形成和发展，人们的注意力转向知识工程，知识工程不同于机器学习那样给计算机输入范例，让它生成出规则，而是直接给计算机输入已被代码化的规则，而计算机是通过使用这些规则来解决某些问题。专家系统就是这种方法所得到的成果，但它有投资大、效果不甚理想等不足。二十世纪80年代，人们又在新的神经网络理论的指导下，重新回到机器学习的方法上，并将其成果应用于处理大型商业数据库。随着在80年代末一个新的术语，它就是数据库中的知识发现，简称KDD(Knowledge discovery in database)。它泛指所有从源数据中发掘模式或联系的方法，人们接受了这个术语，并用KDD来描述整个数据发掘的过程，包括最开始的制定业务目标到最终的结果分析，而用数据挖掘(data mining)来描述使用挖掘算法进行数据挖掘的子过程。但最近人们却逐渐开始使用数据挖掘中有许多工作可以由统计方法来完成，并认为最好的策略是将统计方法与数据挖掘有机的结合起来。

数据仓库技术的发展与数据挖掘有着密切的关系。数据仓库的发展是促进数据挖掘越来越热的原因之一。但是，数据仓库并不是数据挖掘的先决条件，因为有很多数据挖掘可直接从操作数据源中挖掘信息。

五、数据挖掘的应用

数据挖掘的典型应用是在商业领域，但其方法和技术能否应用于其他领域，现在似乎已有突破，如将其应用于医疗领域。总之，有大量数据产生的活动，就应该有应用相关技术的可能。随着技术发展的深入和相关领域知识的渗透，如在天文学、地学、生物学等多方面的潜在应用的可能性是应该存在的。作者在这里提出了问题，但没有论及在不同领域中的具体应用，原因是知识的局限，希望能将数据仓库、数据挖掘等技术应用于科学数据库，从而丰富科学数据库的内容，而将科学数据库的应用推向新的深度。

参 考 文 献

1 Peter Cabena. Discovering Data Mining From Concept to Implementation , IBM, 1997.

2 罗晓沛主编．数据库技术．北京：清华大学出版社，1999.

A Study of Application of Data Mining in Scientific Database

Luo Xiaopei

(Graduate School of Chinese University of Science and Technology, Beijing 100039)

Abstract New application technology in Scientific Database is discussed in this paper, data mining and its application in Scientific Database are also introduced.

Keywords: data mining; Scientific Database.